A Brief History of Protein Sorting Prediction

Nielsen, Henrik; Tsirigos, Konstantinos D.; Brunak, Søren; von Heijne, Gunnar

Published in:
The Protein Journal

DOI:
10.1007/s10930-019-09838-3

Publication date:
2019

Document version
Publisher's PDF, also known as Version of record

The Protein Journal (2019) 38:200–216
https://doi.org/10.1007/s10930-019-09838-3

A Brief History of Protein Sorting Prediction

Henrik Nielsen[1] · **Konstantinos D. Tsirigos**[1] · **Søren Brunak**[1,2] · **Gunnar von Heijne**[3,4]

Published online: 22 May 2019

Abstract
Ever since the signal hypothesis was proposed in 1971, the exact nature of signal peptides has been a focus point of research. The prediction of signal peptides and protein subcellular location from amino acid sequences has been an important problem in bioinformatics since the dawn of this research field, involving many statistical and machine learning technologies. In this review, we provide a historical account of how position-weight matrices, artificial neural networks, hidden Markov models, support vector machines and, lately, deep learning techniques have been used in the attempts to predict where proteins go. Because the secretory pathway was the first one to be studied both experimentally and through bioinformatics, our main focus is on the historical development of prediction methods for signal peptides that target proteins for secretion; prediction methods to identify targeting signals for other cellular compartments are treated in less detail.

Keywords Signal peptides · Protein sorting · Bioinformatics · Prediction

Abbreviations

AA	Amino acid
ANN	Artificial neural network
GO	Gene ontology
HMM	Hidden Markov model
SP	Signal peptide
SVM	Support vector machine
TM	Transmembrane

1 Introduction

The Signal Hypothesis was first proposed by Günter Blobel and David D. Sabatini in a short speculative paper in 1971 [1], where they wrote: "All mRNA's to be translated on bound ribosomes are assumed to have a common feature such as several codons near their 5′ end, not present in mRNA's which are to be translated on free ribosomes. The resulting common sequence of amino acids near the N-terminal of the nascent chains or a modification of it would then be recognized by a factor mediating the binding to the membrane." The postulated "common feature" was first seen in 1972 by Milstein et al. as a larger precursor form of immunoglobulin light chains synthesized in vitro in the absence of rough microsomes [2]. The definitive proof for a cleavable signal peptide (SP) directing protein translocation into the lumen of the ER was published in 1975 by Günter Blobel and Bernhard Dobberstein in their two classic papers describing the in vitro reconstitution of protein translocation [3, 4].

The next obvious question was: what do SPs look like? Are they highly conserved, as suggested by Blobel and Sabatini, or perhaps of more variable sequence? The first data came in 1975 from Edman degradation of an immunoglobulin light chain precursor that had been radiolabeled by [^{3}H]-Leu [5]. The data indicated that the light chain was synthesized with a 20-residue N-terminal extension containing Leu residues in positions 6–8 and 11–13, implying that the SP had a rather hydrophobic character. This was borne out when the full sequences of SPs were starting to be obtained by cDNA sequencing. The first statistical analyses were published in 1979 [6, 7]; they were based, respectively, on collections of 9 and 21, mainly eukaryotic, SPs and noted a

✉ Henrik Nielsen
henni@dtu.dk

1 Department of Health Technology, Section for Bioinformatics, Technical University of Denmark, Kgs. Lyngby, Denmark

2 Faculty of Health and Medical Sciences, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark

3 Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden

4 Science for Life Laboratory, Stockholm University, Solna, Sweden

semi-conserved positively charged N-terminus, a hydrophobic segment, and a C-terminal segment predicted to form a β-strand structure. The 21-sequence collection was analyzed by Garnier et al. in 1980 [8], with similar conclusions.

Sequence patterns around the signal peptidase cleavage site were first discussed in two papers published in early 1983, based on collections of 30 [9] and 78 [10] SPs, respectively; both papers reported that the cleavage site is characterized by having residues with small, uncharged side chains in positions −1 and −3 relative to the cleavage site, an observation referred to as the (−3, −1)-rule.

The first prediction method for SP cleavage sites was described in the paper that introduced the (−3, −1)-rule in 1983 [10]. It was based on a reduced-alphabet weight matrix combined with a rule for narrowing the search region. The weight matrix covered positions −5 to +1 relative to the cleavage site, using only seven different weights at each position, corresponding to groups of amino acids (AAs) with similar characteristics. The weight values were estimated manually rather than calculated from the data. The weight matrix score was only calculated for positions 12–20 counted from the beginning of the h-region—defined as the first quadruplet of AAs with at least three hydrophobic residues—and the cleavage site was assigned to the position with the highest score. This procedure could place the cleavage site correctly in 92% of the data used to estimate it, but measured on a larger data set a few years later, the test performance was only 64% [11]. This serves as a reminder that even a simple method such as a weight matrix can be overfitted, if the underlying data foundation is sparse.

However, this was not the first protein sorting prediction paper. A contender for that title would be the short article by Capaldi and Vanderkooi from 1972 [12], where they show that the proportion of polar residues is different between soluble proteins and integral (at the time called "intrinsic") membrane proteins. At a cutoff of 40% polar residues, roughly half of the membrane proteins were identified with a false positive rate of 6%. This modest success rate was improved by Barrantes, who in papers from 1973 and 1975 [13, 14] developed a linear discriminant function based on two variables: the ratio between charged and hydrophobic AAs, and the average hydrophobicity according to Tanford [15]. While these early works described classification of entire proteins based on overall AA composition, the recognition of individual transmembrane (TM) helices based on AA sequence was pioneered by Kyte and Doolittle [16].

In this review, we describe the early history of protein sorting prediction in detail, while later developments will be mentioned only briefly. The purpose is not to make a complete list of protein sorting prediction software, but to describe those methods that imply significant developments in methodology—one could term them algorithmic paradigm shifts. The main focus will be on SP prediction, with additional sections on TM protein prediction and multi-category protein subcellular location prediction. Specific prediction methods for other sorting signals such as transit peptides for mitochondria and chloroplasts [17, 18], nuclear localization signals [19, 20] and peroxisomal targeting signals [21, 22] will not be mentioned, and are discussed in other, more general, reviews [23–25].

2 Signal Peptide Prediction

Prediction of SPs involves two sub-tasks: discriminating between SPs and non-secretory proteins, and predicting the position of the SP cleavage site. It is important to keep in mind that the presence or absence of an SP is not equal to the question of whether the protein is secreted or not. On the one hand, proteins with SPs may be retained in the membrane or in one of the compartments of the eukaryotic secretory pathway (endoplasmic reticulum, Golgi apparatus or lysosomes); on the other hand, certain proteins are secreted without SPs, especially in bacteria [26], but also in eukaryotes [27].

2.1 A Feature-Based Method

After the reduced weight matrix method appeared in 1983, the first SP predictor was published in 1985 by McGeoch [28] who tested a number of different sequence-derived features to find a combination providing good discrimination between SPs and other sequences. Identification of the precise cleavage site location was not attempted. The two selected features were: length of the uncharged region, and maximal hydrophobicity (on the scale of Kyte and Doolittle [16]) in an 8-AA window. The uncharged region was defined to begin after the last charged AA among the first 11 positions and to end at the next charged residue, while the maximal hydrophobicity was calculated 18 positions downstream from the start of the uncharged region. A non-linear discriminative function, separating the positive and negative examples in the plane defined by these two features, was determined manually.

Originally, this method was based on a very limited data set focusing primarily on virus proteins and immune system proteins, and it could not automatically be transferred to another training set because of the subjective element involved in drawing the separating curve through the two-dimensional feature space. However, the method was later integrated into the multi-category subcellular location predictor PSORT (see Sect. 4), where it is used in combination with von Heijne's 1986 weight matrix (see Sect. 2.2). In PSORT I, the original two features were used for eukaryotic data [29], but for prokaryotic data, the method was retrained using discriminant analysis, and a third feature (net charge of the charged region) was incorporated [30]. For the newer

PSORT II, the method has been further refined for yeast and *Bacillus subtilis*, optimizing not only the coefficients for the features in the discriminant function but also the parameters used to derive the features, i.e., the length of sequence regions scanned for charged or hydrophobic residues, and the hydrophobicity scale [31].

This example nicely illustrates the strengths and weaknesses of rule- and feature-based methods. On the one hand, it is quite transparent how each individual prediction was reached, based on features that are easy to calculate; but on the other hand, the generalization ability is limited. As an example, the rule for finding the start of the "uncharged region" imposes a hard limit on the length of the n-region, so that if an SP has a long n-region containing a charged residue after position 11 (there was one such example in the original data set), the "uncharged region" will not contain the h-region, but only a short arbitrary stretch from the n-region. The feature(s) derived from this will probably be totally out of range for SPs, leaving the method no chance of producing a reasonable answer.

2.2 Weight Matrix Methods

A "real" position-weight matrix, calculated with log-odds scores, was published by von Heijne a few years later [11]. A range of window sizes was tested: Initially, positions −15 to +5 were used, but this could be narrowed to −13 to +2 without loss in performance. Separate matrices were calculated for prokaryotes and eukaryotes.

A common problem for position-weight matrices is that sometimes a certain AA is never observed at a certain position, making it impossible to calculate the logarithm needed. Actually, this situation is only the most extreme instance of the wider problem of sampling errors: The AA distributions are estimated from a limited number of examples, and this tends to overestimate the deviation from a random distribution. The solution is regularization: counteracting the sampling noise by modifying the distribution towards the background. In practice, this is done by adding *pseudocounts* to the observations before calculating the weights [32]. The regularization in this case was done in a rather ad hoc manner: No pseudocounts were added to non-zero counts, while counts of zero were set to one before log-transformation, except in positions −1 and −3 where counts of zero were considered to be significant and were set to 1/N (where N is the number of sequences).

When using the weight matrix for prediction, the weight matrix score was calculated for the first 40 positions of the protein chain, and the cleavage site was assigned to the position with the highest score. Thus, it is an example of a "moving window" method. The maximal weight matrix score in this region was also used for discrimination between SPs and other sequences.

This weight matrix has found extremely wide usage. It was never presented as a mail-server or web-server, but it has been made available as a downloadable program several times [33, 34], it is included in PSORT (see Sect. 4), and it is used together with the McGeoch method [28] in the commercial tool SPScan, which is a part of the widely used Wisconsin Package™ (Genetics Computer Group, GCG). It is also implemented as the "sigcleave" function in the public domain EMBOSS package [35].

In 2004, Hiller et al. made a new set of weight matrices named PrediSi [36] (separate for Gram-negative bacteria, Gram-positive bacteria and eukaryotes). There is no mention of regularization in the article, so apparently the authors did not observe any zero counts. Performance was reported to be close to that of SignalP 2.0 (see Sect. 2.4), but the PrediSi performance was measured by self-consistency only, i.e. without separate training and test sets.

In 2001 [37], Kuo-Chen Chou developed a simple method very similar to a weight matrix, although it was formulated in a different way. Instead of calculating log-odds and summing them over the window, he calculated probabilities and multiplied them (essentially, this is a zero order Markov chain). The probabilities were calculated separately for positive (cleavage site) windows and negative (non-cleavage site) windows and then subtracted to give a discrimination score. The performance was reported to be better than SignalP 1.0 [38] on the same data set, but the comparison was not valid, since it was not the same performance measure that was used. SignalP had reported that the proportion of SP *sequences* where the cleavage site had been correctly placed varied between 68% and 86% (depending on organism group), while Chou reported a 90% correct classification (on all organisms together) of cleavage site versus non-cleavage site *windows*. Essentially, this means that on average every tenth position in a random sequence would be marked as a cleavage site, and there was no indication of how often this result would allow the correct cleavage site to be identified.

Later the same year, Chou modified the method to include the so-called subsite coupling [39], meaning that correlations between selected positions were taken into account. Specifically, the conditional probabilities between positions −3 and −1 and between positions −1 and +1 were included in the calculations. The choice of exactly these positions rested on the fact that they differ most from the background composition, but this in itself is no indication that they are correlated. The modification gave a performance gain of a couple of percentage points. Still the same year, Chou published a version where correlations between all pairs of neighbour positions were used [40] (i.e., a first order Markov chain), which again resulted in a couple of percent better performance. Also in these two papers, the results are compared to the cleavage site performance of SignalP without

acknowledging that the two kinds of percentages represent different performance measures.

The Signal-CF method from 2007 [41] is a further development of the method from Chou's 2001 papers. It is a two-layer system which first determines whether a sequence is an SP or not and subsequently predicts cleavage sites if the sequence is predicted to be an SP. The first layer is based on the so-called pseudo AA composition [42], which essentially means the AA composition augmented by a number of autocorrelation terms that capture some of the sequence order effect by multiplying selected physicochemical parameters of AA pairs separated by a range of different distances. This is used as input to a modified version of the *k* nearest neighbours classifier. The second layer is a weight matrix with "subsite coupling" between positions -3, -1 and $+1$ as described above. Performance was reported to be better than PrediSi, but was not compared to other predictors.

2.3 Early Neural Network Methods

Artificial neural network algorithms (ANNs) learn from data iteratively presented to them by gradually adjusting their weights such that the output values eventually approach desired target values, for example 0.0 or 1.0 representing a dichotomy of whether a protein is secreted or not. Initially the ANNs used in the bioinformatics domain were linear perceptrons without hidden units, such as the ones used by Stormo et al. in 1982 to predict *E. coli* translational initiation sites in nucleotide sequences [43]. While such methods in some cases will outperform rule-based systems, the ANN methodology gained popularity following the reintroduction of the backpropagation algorithm in 1985 by the PDP group [44]. This algorithm allows for training of powerful non-linear models with hidden units that can change their output values significantly in response to small variations in input, for example a change of a single amino acid in a window into the sequence. The backpropagation algorithm was discovered numerous times, however, the pedagogical presentation by Rumelhart et al. [45] quickly led to it becoming widely used, much like the deep learning revolution of today [46]. Non-linear, feed-forward ANNs are quite agile as it is relatively easy to limit overfitting by reducing the number of hidden units [47]. Another feature that has added to the flexibility of the method is that it often is advantageous to combine networks, either in cascades or in a single step [48], a principle allowing for combination of widely different complementary features. This aspect has also had strong impact on the success of ANNs in the protein sorting domain [38].

The first ANN for discrimination between SPs and cytoplasmic proteins was made by Ladunga et al. [49]. Cleavage site prediction was not attempted, and a moving window was not used; instead, the N-terminal part (set to 20 residues after initial testing) of each sequence was used as input. The network was trained with the tiling algorithm, a procedure which builds up the network topology during training, adding as many hidden neurons as necessary to classify all training data correctly [50]. Classifying all training data correctly may sound remarkable, but it leads to virtually guaranteed overfitting—by adding parameters that fit each data point exactly, the network becomes unable to see the forest for the trees. This was reflected in a rather poor test performance when the network was applied to data that had not been part of the training process.

Also in 1991, Arrigo et al. [51] reported that an unsupervised Kohonen network unexpectedly identified the SP region from a small set of human insulin receptor gene data. The Kohonen network, also called a self-organizing feature map, is an example of an unsupervised ANN, where "training" takes place without target values in the training set [50]. The Kohonen network has an input layer and a layer of computational units—the Kohonen nodes. The two layers are fully connected, so that each Kohonen node has a weight vector. The Kohonen nodes are arranged in a way that defines a topological neighbourhood for each node, e.g. a square lattice. When a training example is shown to the network, the Kohonen node whose weight vector is nearest to the input vector is selected. The weight vectors of the selected node and its neighbours within a certain radius are updated, so that they move closer to the input vector by a factor determined by a learning rate. The radius and learning rate decrease during training. In this way, the Kohonen nodes arrange themselves into a pattern that reflects the structure of the input data.

Arrigo et al. trained a network with 30 Kohonen nodes on non-overlapping windows from the cDNA of four human insulin receptor genes. In each sequence, one of the input patterns was extracted as singular in some not very clearly described way; and it turned out that the extracted pattern was wholly or partly within the DNA coding for the SP for a wide range of window sizes. However, it is not clear whether this result has anything to do with SPs at all. Since the approach was not tested on proteins without SPs, the only conclusion to be drawn from this is that the initial part of the reading frame of insulin receptors is in some way peculiar. This might be due to the SP, but it might as well be the effect of correlation between codon bias and intragenic position [52].

Another early ANN was made by Schneider and Wrede [53, 54] who trained a feed-forward ANN to predict SP cleavage sites using moving windows. Instead of sparse encoding, seven physico-chemical properties were used to represent the sequence of AAs. After training networks with a single property at a time, four of them were selected to represent AAs in the final architecture. The training was done with a genetic algorithm rather than backpropagation.

The computations were performed on an extremely small data set derived from *E. coli*: 17 sequences for training and 7 for testing. The final predictor had only 3 of the 7 test cleavage sites correctly placed when assigned by the highest score [54].

After training the predictor, it was used in a "simulated molecular evolution" experiment: a population of 12-aa sequence fragments were subjected to random changes and then selected based on their score for being putative signal sequence cleavage sites according to the ANN. After repeating this for many generations, a number of "optimal" cleavage sites were found, the precise sequence depending on the distance metric used [54]. Remarkably, these all contained Trp, especially at positions −2 and −5, and they had h-regions dominated by Phe. The highest-scoring cleavage site region was subsequently tested in vivo for their ability to promote secretion in an *E. coli* expression system [55]. Indeed, the Phe- and Trp-rich construct (FFFFGWYGWA↓RE) was fully cleavable, but so were the wild type (LAGFATVAQA↓AC) and a "consensus" pattern derived from a simpler, weight matrix-like approach (VVIMSASAMA↓AC).

Although this whole process is based on statistics from only 24 sequences, the result raises an interesting point: When using a linear method, the optimal example looks like a consensus of the training examples; but for a non-linear method, this is not necessarily the case. It is remarkable that the highest-scoring examples according to the ANN are very rich in otherwise rare AAs. So, is there any reason to expect that the non-linearly optimized "FFFFGWYGWA↓RE" is a more efficient cleavage site than the linearly optimized "VVIMSASAMA↓AC"? Probably not. Even if we assume that the peculiar residues are not just an effect of sampling error, the highest ANN score is found in a region of sequence space not covered by the training data, implying that the network score here is an *extrapolation* rather than an *interpolation*. And since ANNs do not contain any physicochemical model of how scores should vary with the input, but simply fit a non-linear function to the examples, a good generalization in interpolation does not necessarily mean a good generalization in extrapolation. The more non-linear the fitted function is, the less we can assume about how it should continue outside the region of the fitted data.

The Schneider and Wrede 1994 paper [54] was harshly criticized in a comment by Darius and Rojas [56] who, among other points, wrote: "The term "quality" for the value of the fitted function gives the impression that some biological significance is associated with values of the fitted function strictly between 0 and 1, but there is no justification for this kind of interpretation and finding the point where the fit achieves its maximum does not make sense."

2.4 SignalP and Related Neural Network Applications

SignalP 1.0 [38, 57] was in 1996 the first machine learning based SP prediction method to be available online as a web-server. SignalP used a combination of two different ANNs with moving windows: one trained to recognize all positions within the SPs, and one trained to recognize the cleavage site specifically. The outputs of these two networks were termed S-score and C-score, respectively. These were then combined into the Y-score, which was a function of the C-score and the slope of the S-score, used for predicting the location of the cleavage site. This way of combining two ANNs was inspired by the intron splice site predictor NetGene from 1991 [48].

SignalP 2.0 from 1999 [58] added a hidden Markov model (HMM, see Sect. 2.5) which made it possible to distinguish cleaved SPs from uncleaved signal anchors, while SignalP 3.0 from 2004 [59] introduced the D-score (the average of maximal Y-score and mean S-score) as a better discriminator between SPs and other sequences. SignalP 4.0 from 2011 [60] brought a new definition of the negative data: instead of just soluble intracellular proteins and signal anchors, it now included all TM proteins that had a TM helix within the first 70 positions and therefore could be mistaken for SPs. This drastically reduced the number of false positives produced by TM proteins. Unfortunately, SignalP 4.0 also had a lower sensitivity than SignalP 3.0 which led to many complaints from users whose favourite SPs were suddenly no longer positively predicted. Therefore, SignalP was in 2012 updated to version 4.1 with an option to choose an alternative threshold that reproduced the sensitivity of SignalP 3.0 [61].

In prokaryotes, there are several types of SPs. SignalP versions 1–4 were only able to predict the "standard" type of SP, which is transported through the Sec translocon and cleaved by signal peptidase I (also known as leader peptidase). However, there are also specialized SPs of prokaryotic lipoproteins which are cleaved by signal peptidase II (also known as lipoprotein signal peptidase); these have a different cleavage site motif with a 100% conserved cysteine in the +1 position [62]. In addition, there are SPs that direct their proteins through the Tat translocon; these have a characteristic twin-Arginine motif in the n-region [63] and are typically longer and less hydrophobic than Sec SPs [64]. In our group, separate ANNs were trained to predict such SPs, constituting the cores of the prediction methods LipoP from 2003 [65] and TatP from 2005 [66], respectively.

In 2003, an Italian group published SPEPlip [67], an ANN-based method very similar in architecture to SignalP. It was combined with a simple PROSITE pattern [68] which made it possible to distinguish between "standard" SPs cleaved by signal peptidase I and lipoprotein SPs cleaved by signal peptidase II.

The ANNs mentioned so far have generally had at most one hidden layer of computational neurons. The original backpropagation algorithm did not work well for deep networks with many layers, as the error made at the final output layer is not easy to use as a precise measure further up in the layered structure for adjusting the weights. In a deep ANN the many layers can filter and reorder features in very powerful ways. That is generally not possible using a single hidden layer unless one uses a huge number of units that most often will lead to overfitting. Deep ANN architectures are capable of carrying out sophisticated feature engineering as opposed to just establishing a simple decision boundary in the feature space resulting from the more moderate feature engineering achievable by the input-to-hidden transformation (when the hidden layer is of moderate size). The newer deep learning techniques solve these problems while keeping the number of adjustable parameters down. The techniques can be applied to feed-forward networks with many layers, but also to recurrent networks with loops that can memorize features that correlate with a desired output [69–72].

Deep learning in SP prediction was introduced in 2017 by the method DeepSig [73] (from the same group that published SPEPlip). It is based on convolutional ANNs, which can be described as a set of moving windows that look at small portions of the input sequence at a time. In DeepSig, there are three consecutive combinations of a convolutional layer feeding into an average pooling layer. This is followed by a so-called Taylor decomposition, a layer that estimates the relevance of each position in the input sequence for the classification of the sequence as an SP or not. Finally, the cleavage site is assigned by a grammar-restrained conditional random field, a probabilistic model which resembles an HMM in having a grammatical structure defining, in this case, the three regions of the SP. DeepSig was trained on the data from SignalP 4.0 and was reported to outperform it in most cases.

The recently released SignalP 5.0 [74] is based on deep ANNs of the recurrent type, where information flows not only from input to output, but also between the hidden units. The recurrent architecture of SignalP 5.0 makes it possible to abandon the moving windows, which defined the C- and S-scores in earlier versions of SignalP. Instead, the so-called long short-term memory networks can take a sequence of varying length as input and, if necessary, remember features from the beginning of the sequence while classifying positions further downstream [75]. The output from the long short-term memory layer is passed on to a conditional random field specifying that a cleavage site can only follow after an SP position and must necessarily be followed by a mature protein position. In this way, post-processing in the form of calculating Y- and D-scores becomes unnecessary.

Another innovation in SignalP 5.0 is that it now can predict SPs using the Tat pathway and lipoprotein SPs cleaved by signal peptidase II, meaning that a user no longer has to consult three different predictors in order to get a prediction of which type a prokaryotic SP belongs to.

2.5 Hidden Markov Models

In SignalP 2 and 3 [58, 59], an HMM predicted SPs independently of the ANN. This HMM was not of the profile type that has found wide usage in databases of protein families such as Pfam [76]; instead, it reflected the usual description of SPs as consisting of n-, h-, and c-regions. The n-region and the h-region were modeled by common AA distributions; only around the cleavage site were single positions modeled separately. Instead of C-scores and S-scores, the HMMs provided probabilities of the three regions and of the cleavage site.

The original rationale for employing an HMM in SignalP 2 was to facilitate discrimination between SPs and signal anchors (uncleaved transmembrane helices close to the N-terminus). The distinction between SPs and signal anchors is not merely a question of having a cleavage site or not; signal anchors typically have hydrophobic regions longer than those of SPs. Interestingly, experiments have shown that it is possible to convert a cleavable SP to a signal anchor merely by lengthening the h-region [77, 78]. Our idea was that an HMM better than an ANN would be able to model this length difference. However, when constructing SignalP 4 [60] and retraining the HMMs on the new data set, we found them to be inferior in performance to the ANNs, thereby disproving our original idea. Apparently, ANNs with sufficiently large input windows are able to discriminate between short and long hydrophobic regions.

It is not impossible to construct a profile HMM that recognizes SPs; this was done by Zhang & Wood in 2003 [79]. However, its performance did not quite match that of the HMM module in SignalP 2.0.

The HMM-based Phobius TM topology prediction method from 2004 also includes an SP model [80]. This carries two advantages for TM prediction: first, false positive predictions of TM helices in SP regions are avoided; second, the topology of TM proteins carrying SPs is constrained by the fact that the N-terminus of the mature protein must be on the non-cytoplasmic side. The SP model in Phobius very closely resembles the one used in SignalP 2 and 3.

Similar to TatP [66] and LipoP [65], specialized prediction methods based on HMMs have also been presented. PRED-TAT [81] aims at discriminating between Tat and Sec-translocated SPs, as well as predicting their cleavage sites. PRED-LIPO [82] predicts presence of Sec/SPI SPs and Sec/SPII SPs in Gram-positive bacteria and can discriminate them from cytoplasmic and N-terminal TM proteins. Finally, PRED-SIGNAL [83] was the first computational method

that specifically predicts SPs of archaeal origin and their cleavage sites, using an HMM approach.

2.6 Support Vector Machine Applications

Unlike ANNs and HMMs, the third major machine learning algorithm, support vector machines (SVMs), has not played a big role in SP prediction. This is in contrast to the situation in prediction of subcellular location based on AA composition, where SVMs have been very important (see Sect. 4.1). One exception was made by Vert in 2002 [84], who trained an SVM for SP cleavage sites using a new class of kernels for strings. He used a −8 to +2 window and discriminated between windows with and without a cleavage site, and showed that the SVM was superior to a retrained weight matrix on the same data set. However, a comparison to an ANN was not made.

The year after, Cai et al. [85] published an SVM for predicting SP cleavage sites using sparse encoding of the inputs and a polynomial kernel. The resulting performance was not compared to anything, but it was slightly worse than the "subsite coupling" method by Chou [39] on the same dataset. In 2005, Wang et al. [86] tackled the same problem with a string kernel. They did an extensive comparison to a retrained weight matrix using the same dataset. For small windows (−8 to +2), the SVM outperformed the weight matrix, but for larger windows (−13 to +2 or larger) the advantage of the SVM disappeared.

Another SVM method is the TM topology predictor MEMSAT-SVM [87], which also predicts SPs. MEMSAT-SVM is built from five binary window-based classifiers, of which one is SP/non-SP. They are trained using the traditional polynomial or radial basis function kernels rather than a string kernel. MEMSAT-SVM is especially interesting because it can be compared to the ANN-based MEMSAT3 [88] which was published 2 years earlier. MEMSAT-SVM performed better than MEMSAT3 on almost all parameters.

Cleavage sites are not explicitly recognized by the windows of the MEMSAT methods, and the cleavage site performance is not reported in the papers. When testing version 4.0 of SignalP [60], we benchmarked MEMSAT3 and MEMSAT-SVM and found that both of them had cleavage site precision and recall values close to zero. Regarding discrimination between SPs and non-SPs, we could confirm that MEMSAT-SVM was better than MEMSAT3, but it was still not among the best performing methods.

2.7 Homology-Based Methods

Signal-3L from 2007 [89] is a further development of the two-layer Signal-CF method [41] (see Sect. 2.2). It is mentioned in this section because it adds a third layer where alignment is used for improving the cleavage site prediction. The second layer suggests a number of cleavage sites, and then global pairwise alignment to a database of known SPs is used for selecting the best candidate among them. Performance was reported to be better than PrediSi [36], but was not compared to other predictors.

Signal-BLAST from 2008 [90] is a much simpler prediction method that runs BLAST [91] against a pre-constructed reference database of SPs, and, if it finds a hit with high similarity, it assigns the cleavage site position based on the homologous protein. This approach works very well if there are annotated close homologues in the database, but the drawback of this approach is that its performance solely depends on sequence similarities that can be detected by the BLASTP algorithm. The authors do not report on the tool's performance when low or no sequence similarities are found. In our hands [74], Signal-BLAST did not perform well when no hits to its reference database were found, since it does not have a fallback strategy for these cases.

In 2017, Signal-3L was updated to version 2.0 [92] with a major modification of the architecture of the method. The first layer is now an SVM taking its input from PSI-BLAST [91] profiles, predicted secondary structure, predicted disorder, and selected physicochemical parameters, while the second layer searches the Conserved Domain Database [93] for functional domains in order to distinguish between SPs and TM helices. The third layer then corresponds to the second *and* third layer of the original Signal-3L. The performance was in some cases reported to be better than that of SignalP 4.1 [60], although SignalP 4.1 always had the lowest false positive rates. In the SignalP 5.0 benchmarks [74], however, Signal-3L 2.0 was not better than SignalP 4.1.

When using homology to predict SPs and their cleavage sites, it should be noted that SPs (and other N-terminal sorting signals) are actually *less* conserved than the mature regions of the proteins [94]. Therefore, it may be beneficial to search a database of entire proteins instead of a database of SPs.

3 Transmembrane Protein Prediction

TM proteins constitute one of the most well-studied categories of membrane proteins. In numbers, they make up roughly 30% of the total number of proteins in a fully-sequenced organism, and their roles are diverse and important to the life of the cells [95, 96]. An important obstacle in the study of TM proteins is the difficulty in the determination of their 3D-structure, owing mainly to their hydrophobic nature [97]. The emergence of automated, computational methods that provide the researchers with a topological model of TM proteins has been very important to the field. These models inform about the number and position of the TM segments, alongside with the orientation with regards

to the membrane. A major challenge in obtaining successful topology predictions is the exact same hydrophobic nature, which results in erroneous assignment of N-terminal TM segments as SPs and vice versa [98, 99].

As mentioned, Kyte and Doolittle in 1982 initiated the prediction of TM helices in a paper that was concerned with displaying the hydropathic character of proteins in general [16]. To this end, they developed a novel hydropathy index based on water/vapour transfer energies, buried/exposed propensities, and certain manual adjustments, the latter being described as "the result of personal bias and heated discussion between the authors". Their program, SOAP, calculated the average hydropathy value of overlapping *k*-residue segments. While they found $k=9$ to give the best correlation with buried and exposed stretches of globular proteins, $k=19$ yielded the best discrimination between TM segments and hydrophobic stretches of globular proteins.

Later the same year, Argos et al. [100] published a method for predicting the structure of TM proteins. Instead of settling for one hydrophobicity scale, they investigated nine different properties of AAs and used a fitting procedure to the proposed structure of bacteriorhodopsin to adjust the weights for each property. Five of the nine properties were eventually selected. Instead of calculating averages within a fixed length window, a smoothing procedure was used. While a good agreement with the bacteriorhodopsin structure was achieved, the method was not very good at discriminating between TM segments and globular proteins.

The ALOM method from 1985 [101] was very similar to SOAP, but based on a larger data set. Four different hydrophobicity scales were tested, and the Kyte-Doolittle hydropathy was eventually chosen. The authors found a 17-residue window to give the best discrimination between integral and peripheral membrane proteins, and devised an additional procedure for assigning the precise boundaries between TM helices and loops. ALOM was later incorporated as a feature in the PSORT prediction method (see Sect. 4).

These early topology prediction methods were based on the amino acids' hydrophobicity as a way to detect potential TM regions in the sequence, but were unable to inform regarding their orientation. This changed with the observation that positively-charged residues are more frequently found on the cell's 'inside' (cytoplasmic loops), an observation widely known as the 'positive-inside rule' [102, 103]. This finding was implemented in the TopPred algorithm from 1992 [104, 105], where, for the first time, the software could decide whether a given region is cytoplasmic, extracellular or TM.

In 1994, the MEMSAT algorithm [106, 107] used statistical tables compiled from well-characterized membrane protein data and, by combining dynamic programming and propensity scales, produced the best overall topology. In the following years, more methods were made available to the public, based on statistical analysis of amino acid preferences and hydrophobicity, like PRED-TMR [108] and the more recent SCAMPI [109], which has been updated in 2016 [110].

The use of HMMs for the topology prediction task was initially introduced in the TMHMM [95, 111] and HMMTOP [112, 113] methods in 1998. Some years after these first HMM-based attempts, given that SPs are often falsely predicted as TM segments because of their high hydrophobicity, better-scoring methods that simultaneously predict the topology of the protein and the presence of an SP were developed, beginning with Phobius in 1994 [80]. Later developments along these lines include PolyPhobius (using evolutionary information [114]), Philius [115], MEMSAT3 [88], MEMSAT-SVM [87] and SPOCTOPUS [116].

A key improvement in the topology prediction field was the inclusion of evolutionary information in the prediction process, in the form of multiple sequence alignments, also known as profiles. The early algorithms used only a single sequence as input; however, as the sequence databases were growing with time, researchers started to exploit the availability of data. In 1993, it had been shown that profiles do improve protein secondary structure prediction [117]. The methods TMAP from 1994 [118] and PHDhtm from 1995 [119] were the first to use the evolutionary information in topology prediction. This step, as was later shown in a comparative study [120], indeed improves the accuracy and has since many years become a standard step during the development of sequence-based prediction algorithms.

PHDhtm [119] was the first topology prediction method that incorporated ANNs in the prediction process for TM proteins. By using profiles, it creates a consensus prediction for the target sequence and then finds the topology of the protein using the "positive-inside rule". Similarly, methods that also use evolutionary information, like PRO/PRODIV-TMHMM [120] and OCTOPUS [121] were created. The latter is a combination of ANNs, that predict inside/outside and membrane/non-membrane residue preferences and an HMM which is then used to calculate the final topology.

Other machine learning methods that have been used in topology predictors are Support Vector Machines (SVMs) and Dynamic Bayesian Networks (DBNs) that are found in MEMSAT-SVM [87] and Philius [115], respectively. Finally, consensus-based approaches, like CoPreTHi [122], TOPCONS [123], MetaTM [124] and CCTOP [125], which combine the outputs from several predictors into a consensus output using dynamic programming, have been quite successful.

Prediction methods that include both models for SPs and TM segments [80, 87, 88, 114–116] are more useful for proteome-wide analyses. The updated version of the TOPCONS consensus topology prediction method, TOPCONS2 [126],

can also account for the presence of an SP, thus is ideal for large scale predictions.

Numerous methods that aim at topology prediction of β-barrel TM proteins also exist. These include methods based on hydrophobicity analysis [127], statistical preferences of amino acids [128], remote homology detection [129], HMMs [130–134], SVMs combined with HMMs [135, 136], feed-forward ANNs [137, 138] and radial basis function ANNs [139]. PRED-TMBB2 [134] is, to our knowledge, the only approach available as a web-server that incorporates SP prediction in the topology prediction. This is an important feature since bacterial β-barrel proteins should have an SP that guides them through the inner membrane and towards the outer membrane of the cell. More detail on prediction of both α-helix and β-barrel TM proteins is found in a recent review [97].

4 Multi-category Location Predictors

Obviously, the presence or absence of an SP or one or more TM segments is not the whole story about the subcellular location of a protein. The typical user will want to know not only whether certain sorting signals are present, but exactly where in the cell the protein goes. Several predictors have attempted to deliver this service, the first being PSORT from 1991 [29, 30]. This was an integrated expert system of several prediction methods, using both sorting signals and global properties. Some of the components were developed within the PSORT group, others were implementations of methods published elsewhere, including selected PROSITE patterns [68]. PSORT was the first publicly available system that showed this degree of integration, and it included predictions for locations that no other available methods provided at that time, e.g., nuclear and peroxisomal targeting.

All the constituent predictors provided feature values, which were then integrated to produce a final prediction. In the original version, PSORT I, the integration was done in the style of a conventional knowledge base using a collection of "if–then" rules. This makes it very difficult to adjust the rules according to information from new data sets; so in order to be able to incorporate new data on a regular basis, the newer PSORT II version used quantitative machine learning techniques, such as probabilistic decision trees and the *k* nearest neighbours classifier to integrate scores from all the features [140, 141].

4.1 Amino Acid Composition-Based Methods

In addition to the recognition of sorting signals, prediction of protein sorting can exploit the fact that proteins of different subcellular compartments differ in global properties, reflected in the AA composition. While the signal prediction methods are probably closer to mimicking the information processing in the cell, methods based on global properties can complement imperfect signal-based methods, especially on incomplete sequences. Specifically, a composition-based method for recognizing extracellular proteins can be used without knowledge of the N-terminus, and could give correct predictions for, e.g., protein fragments or genomic sequences with erroneous assignment of start codons. One drawback is that such methods will not be able to distinguish between closely related proteins that differ in the presence or absence of a sorting signal.

As mentioned in the introduction, this approach constituted the very beginning of protein sorting prediction in the attempts by Capaldi and Vanderkooi and Barrantes [12–14] to recognize integral membrane proteins. In 1994, Nakashima and Nishikawa [142] reestablished this line of research by using simple odds-ratio statistics to discriminate between soluble intracellular and extracellular proteins on the basis of AA composition and AA-pair frequencies. Including AA pairs (separated by up to four positions) improved performance by 8% relative to AA composition alone.

In 1997, Cedano et al. [143] extended the number of possible locations to five: intracellular, extracellular, transmembrane, membrane-anchored, and nuclear; and used the so-called Mahalanobis distance to discriminate. This metric takes interactions between AAs into account (note: not interactions between positions; the input is only the 20 AA frequencies) and is therefore able to handle non-linear mappings in the 20-dimensional space defined by the AA composition. Their algorithm, named ProtLock, was for a time available as a downloadable program. This approach was refined in three subsequent papers by Chou and Elrod [144–146], who used a modified version of the Mahalanobis distance, where an extra term compensated for differences in size between the categories.

The NNPSL method by Reinhardt and Hubbard from 1998 [147] used ANNs trained on overall AA composition to predict location. They discriminated between three bacterial compartments (cytoplasmic, periplasmic, and extracellular) and four animal/fungal compartments (cytoplasmic, extracellular, mitochondrial, and nuclear). Interestingly, plant proteins were found to be very poorly predicted, and were not included in the final method. The NNPSL dataset was subsequently used by others employing different machine learning techniques, notably Yuan in 1999 using Markov chains [148] and Hua and Sun in 2001 using SVMs in their method named SubLoc [149].

One rather disturbing aspect of these early composition-based methods is their lack of proper homology reduction of the data. If the test set contains sequences that are very closely related to sequences in the training set, these proteins will also be close to each other in AA composition

space, and prediction performance will be overestimated. To estimate a true generalization performance on new unrelated sequences, the dataset should be reduced or partitioned to avoid homology between training and test, and the threshold for when two proteins are too closely related should be set at a value where the problem cannot be solved by alignment alone. In the field of protein structure prediction, methods for determining the threshold and carrying out the reduction were published in the early 1990s [150, 151], and concerning SP prediction, the choice of threshold was discussed in detail in a 1996 paper [152], but in AA composition-based methods, it was apparently ignored for another decade. The NNPSL dataset was homology reduced, but only down to 90% identity [147], while Chou and Elrod only excluded proteins with the same *name* from the data set [144–146].

Better homology reduction was introduced to this subfield in 2005 and 2006 by three SVM-based methods, LocTree [153], CELLO [154] and BaCelLo [155]. They all supplement the total AA composition (and, in the case of CELLO, also AA pair composition) by AA composition in parts of the sequence. While CELLO divided the sequence into a number of subsequences of equal length, BaCelLo used a set of N- and C-terminal windows of fixed lengths, and LocTree calculated AA composition for three predicted secondary structure states separately. Both BaCelLo and LocTree searched a sequence database to create a profile and calculated AA composition in these profiles rather than the query sequence itself. The CELLO authors made a thorough examination of the relationship between alignment and machine learning predictions and reported that above a limit of 30% identity, alignment performed better than the SVM-based prediction system. Similarly, the BaCelLo authors reported that the prediction of subcellular localization in the NNPSL dataset could be carried out with a BLAST search [91], where the localization of each protein was simply predicted to be that of the closest homologue within the dataset. The performance of this simple procedure was actually better than that of the machine learning-based methods NNPSL and SubLoc and at the same level as two newer methods (LOCSVMPSI [156] and ESLpred [157]).

Why does the AA composition approach work to some degree, if it is not able to detect the sorting signals? It is no mystery that discrimination of TM versus soluble proteins is possible, since the strong hydrophobicity of the TM helices influences the AA composition; and the discrimination of inner versus outer membrane TM proteins should also be quite easy, since these are generally α-helix versus β-sheet proteins, respectively. It is more surprising that discrimination between soluble proteins of different compartments by AA composition is possible. A plausible explanation is that the protein surfaces reflect the chemical properties (acidity, ion concentrations, etc.) of their compartments. Andrade et al. [158] found that the signal in the total AA composition, which makes it possible to identify the subcellular location, is due almost entirely to surface residues.

4.2 Homology-Based Methods

Arguably, the simplest approach to subcellular location prediction is the BLAST search described in the previous subsection: assign the subcellular location of the best hit in a database of annotated examples. This is based on the assumption that proteins tend to stay in the same compartment over the course of evolution, which seems to be the case judging from an extensive analysis of sequence conservation in relation to subcellular location [159].

Imai and Nakai in 2010 [160] showed that this approach was superior to three at the time well established predictors (CELLO 2.5 [154], MultiLoc2 [161] and WoLF PSORT [162]) if the dataset was not homology reduced, and it performed on a par with the predictors if the dataset was homology reduced to 30% identity. This result was used by the authors of the LocTree3 method in 2014 [163]: It simply outputs the location of the best BLAST hit in an annotated database if that hit has an E-value better than a certain cutoff, and reverts to its predecessor, the SVM-based LocTree2 [164], otherwise. The bacteria-specific predictor PSORTb 3.0 [165] uses a similar combination of approaches.

There are other ways to use homology information than this direct transfer of homologue location annotation. One approach is to calculate a phylogenetic profile for each protein—a specification of the pattern of occurrence of matches to that protein among a set of organisms with sequenced genomes. This was pioneered by Marcotte et al. [166]. Another approach is to search for conserved domains or motifs that are characteristic of specific locations [167].

It is also possible to use other parts of the homologue annotations than the subcellular location information. Several predictors use Gene Ontology (GO) terms [168] of the retrieved homologues as inputs for their methods, including the GOASVM and mGOASVM predictors [169, 170] and the iLoc family of predictors [171–177]. The GO terms may contain a richer source of information, but they also frequently include terms that are themselves predicted, potentially leading to a situation of circular reasoning, especially if GO-based predictors are used for assigning new GO terms. Other approaches in this direction are taken by the PA-SUB predictor [178] which looks at the occurrence of certain key words and phrases in the UniProt entries of the retrieved homologues, and the SherLoc predictor [179, 180] which does text mining of abstracts linked from the UniProt entries.

There are two advantages to using AA composition-based or homology-based methods. First, they can be used also for those compartments where the actual sorting signals are not known, or are too poorly characterized to support a proper

signal-based prediction method. Second, they may work for sequences that are fragments from which the actual sorting signal may be missing or for amino acid sequences derived from genomic or metagenomic sequence where the start codon of the protein has not been correctly predicted, thus obscuring any N-terminal sorting signals. On the downside, AA composition-based or homology-based methods do not provide the same degree of insight into the information processing in the cell since they typically ignore which parts of the sequence are actually important for sorting. Another drawback is that such methods will not be able to distinguish between very closely related proteins that differ in the presence or absence of a sorting signal, and they will not be able to predict the effects of small mutations that destroy or create a sorting signal.

4.3 Integrated Methods

As mentioned, PSORT I was in 1991 the first integrated method for protein subcellular location prediction. It was succeeded by PSORT II in 1996 [140, 141] and later by PSORTb for bacterial proteins [165, 181, 182] and WoLF PSORT for eukaryotic proteins [162]. All these methods are based on feature predictors that predict e.g. SPs or TM helices, and classification systems that integrate the output of the feature predictors. A homology component is, as mentioned, also part of PSORTb 3.0 [165].

A similar approach was taken by MultiLoc in 2006 [183] which integrated SVM-based prediction of N-terminal sorting signals with SVM-based prediction based on AA composition and a database of sorting-relevant motifs such as nuclear localization signals. The integration was done by another layer of SVMs. MultiLoc2 from 2009 [161] additionally incorporated phylogenetic profiles and GO terms of retrieved homologues.

YLoc from 2010 [184, 185] used a different technology, the Naïve Bayes classifier, to select between a very large number of simple features and integrate the selected features. Naïve Bayes is a linear method that is often outperformed by ANNs, HMMs, or SVMs. The advantage, however, is that it not only provides a prediction, but also a *reason* for the prediction in the form of a list of the features that led to the prediction in each particular case. YLoc can optionally include GO terms in the prediction.

LocTree2 from 2012 [164] is a system of SVMs arranged in a hierarchy or decision tree. Each decision is made by an SVM using a *profile kernel*, a kind of string kernel that calculates the frequencies of short motifs in a sequence profile made by PSI-BLAST [91].

Lastly, deep learning has also entered the multi-location prediction field in the form of DeepLoc from 2017 [186]. DeepLoc uses a combination of convolutional and recurrent ANNs together with a so-called attention layer which assigns a weight to every position in the sequence. In this way, the user gets an indication of which parts of each input sequence were important for the prediction. DeepLoc does not use any annotation from homologues, but still its performance was shown to be superior to seven other methods including GO-based ones like MultiLoc2 [161] and iLoc-Euk [171].

5 Discussion

The first attempts to predict SPs in protein sequences were made more than 35 years ago, based on very simple statistics. Since then, the field has progressed in steps with methodological developments in the wider area of bioinformatics, such as the use of weight matrices, ANNs, HMMs, SVMs, and, more recently, deep and recurrent ANNs. In turn, these increasingly "data-hungry" approaches have been made possible by the revolution in high-throughput DNA sequencing that we have witnessed over the past couple of decades. From a historical perspective, this can stand as a nice example of how developments in computer science and wet-lab molecular biology have reinforced each other in creating the vast field of sequence-based bioinformatics that we see today.

It is difficult to estimate the full impact of SP prediction methods on biology, but it is abundantly clear that they have played an important role in proteomics research, genome annotation, identification of potential drug targets, and in a multitude of cases where the knowledge that a particular protein is secreted or membrane-anchored has been critical for understanding its function.

In the age when a rapidly growing number of genomes have been sequenced, experimentally confirmed annotations of subcellular location, molecular function, post-translational modifications etc. do not grow at nearly the same pace. Consequently, the islands of experimental annotations are increasingly far apart in the expanding sea of sequence data. This means that homology-based methods, which depend critically on the quality of the annotations they use for prediction, have an increasingly sparse basis of high quality data. It also means that machine learning methods should find a way to utilize the information inherent in unannotated data.

As our knowledge of cell architecture and compartmentalization improves, the need for new and even better methods to predict subcellular localization of proteins will remain, also given the revolution in single cell technologies. In particular, there is still room for major improvements in signal-based multi-category location methods that can sort proteins between multiple cellular locations with high reliability through modeling of the actual sorting signals, instead of relying on AA composition or homology. While in many

ways mature, the field still holds interesting challenges for the bioinformatician.

Acknowledgements SB would like to acknowledge support from the Novo Nordisk Foundation (grant NNF14CC0001).

References

1. Blobel G, Sabatini DD (1971) Ribosome-membrane interaction in eukaryotic cells. In: Manson LA (ed) Biomembranes. Plenum Press, New York, pp 193–195
2. Milstein C, Brownlee GG, Harrison TM, Mathews MB (1972) A possible precursor of immunoglobulin light chains. Nat New Biol 239:117–120. https://doi.org/10.1038/newbio239117a0
3. Blobel G, Dobberstein B (1975) Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. J Cell Biol 67:835–851. https://doi.org/10.1083/jcb.67.3.835
4. Blobel G, Dobberstein B (1975) Transfer of proteins across membranes. II. Reconstitution of functional rough microsomes from heterologous components. J Cell Biol 67:852–862. https://doi.org/10.1083/jcb.67.3.852
5. Schechter I, McKean DJ, Guyer R, Terry W (1975) Partial amino acid sequence of the precursor of immunoglobulin light chain programmed by messenger RNA in vitro. Science 188:160–162. https://doi.org/10.1126/science.803715
6. von Heijne G, Blomberg C (1979) Trans-membrane translocation of proteins. Eur J Biochem 97:175–181. https://doi.org/10.1111/j.1432-1033.1979.tb13100.x
7. Austen BM (1979) Predicted secondary structures of amino-terminal extension sequences of secreted proteins. FEBS Lett 103:308–313. https://doi.org/10.1016/0014-5793(79)81351-4
8. Garnier J, Gaye P, Mercier J-C, Robson B (1980) Structural properties of signal peptides and their membrane insertion. Biochimie 62:231–239. https://doi.org/10.1016/S0300-9084(80)80397-X
9. Perlman D, Halvorson HO (1983) A putative signal peptidase recognition site and sequence in eukaryotic and prokaryotic signal peptides. J Mol Biol 167:391–409. https://doi.org/10.1016/S0022-2836(83)80341-6
10. von Heijne G (1983) Patterns of amino acids near signal-sequence cleavage sites. Eur J Biochem 133:17–21. https://doi.org/10.1111/j.1432-1033.1983.tb07424.x
11. von Heijne G (1986) A new method for predicting signal sequence cleavage sites. Nucleic Acids Res 14:4683–4690. https://doi.org/10.1093/nar/14.11.4683
12. Capaldi RA, Vanderkooi G (1972) The low polarity of many membrane proteins. Proc Natl Acad Sci USA 69:930–932. https://doi.org/10.1073/pnas.69.4.930
13. Barrantes FJ (1973) A comparative study of several membrane proteins from the nervous system. Biochem Biophys Res Commun 54:395–402. https://doi.org/10.1016/0006-291X(73)90935-2
14. Barrantes FJ (1975) The nicotinic cholinergic receptor: different compositions evidenced by statistical analysis. Biochem Biophys Res Commun 62:407–414. https://doi.org/10.1016/S0006-291X(75)80153-7
15. Tanford C (1962) Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. J Am Chem Soc 84:4240–4247. https://doi.org/10.1021/ja00881a009
16. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132. https://doi.org/10.1016/0022-2836(82)90515-0
17. Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300:1005–1016. https://doi.org/10.1006/jmbi.2000.3903
18. Savojardo C, Martelli PL, Fariselli P, Casadio R (2015) TPpred3 detects and discriminates mitochondrial and chloroplastic targeting peptides in eukaryotic proteins. Bioinformatics 31:3269–3275. https://doi.org/10.1093/bioinformatics/btv367
19. Cokol M, Nair R, Rost B (2000) Finding nuclear localization signals. EMBO Rep 1:411–415. https://doi.org/10.1093/embo-reports/kvd092
20. Brameier M, Krings A, MacCallum RM (2007) NucPred—predicting nuclear localization of proteins. Bioinformatics 23:1159–1160. https://doi.org/10.1093/bioinformatics/btm066
21. Emanuelsson O, Elofsson A, von Heijne G, Cristóbal S (2003) In silico prediction of the peroxisomal proteome in fungi, plants and animals. J Mol Biol 330:443–456. https://doi.org/10.1016/S0022-2836(03)00553-9
22. Neuberger G, Maurer-Stroh S, Eisenhaber B et al (2003) Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. J Mol Biol 328:581–592. https://doi.org/10.1016/S0022-2836(03)00319-X
23. Emanuelsson O (2002) Predicting protein subcellular localisation from amino acid sequence information. Brief Bioinform 3:361–376. https://doi.org/10.1093/bib/3.4.361
24. Nakai K, Horton P (2007) Computational prediction of subcellular localization. In: Giezen M (ed) Protein Targeting Protocols. Humana Press, New York, pp 429–466
25. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2:953–971
26. Desvaux M, Hébraud M, Talon R, Henderson IR (2009) Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue. Trends Microbiol 17:139–145. https://doi.org/10.1016/j.tim.2009.01.004
27. Duitman EH, Orinska Z, Bulfone-Paus S (2011) Mechanisms of cytokine secretion: a portfolio of distinct pathways allows flexibility in cytokine activity. Eur J Cell Biol 90:476–483. https://doi.org/10.1016/j.ejcb.2011.01.010
28. McGeoch DJ (1985) On the predictive recognition of signal peptide sequences. Virus Res 3:271–286. https://doi.org/10.1016/0168-1702(85)90051-6
29. Nakai K, Kanehisa M (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics 14:897–911. https://doi.org/10.1016/S0888-7543(05)80111-9
30. Nakai K, Kanehisa M (1991) Expert system for predicting protein localization sites in gram-negative bacteria. Proteins Struct Funct Bioinform 11:95–110. https://doi.org/10.1002/prot.340110203
31. Nakai K (1996) Refinement of the prediction methods of signal peptides for the genome analyses of *Saccharomyces cerevisiae* and *Bacillus subtilis*. Genome Inform 7:72–81
32. Henikoff JG, Henikoff S (1996) Using substitution probabilities to improve position-specific scoring matrices. Bioinformatics 12:135–143. https://doi.org/10.1093/bioinformatics/12.2.135
33. Folz RJ, Gordon JI (1987) Computer-assisted predictions of signal peptidase processing sites. Biochem Biophys Res Commun 146:870–877. https://doi.org/10.1016/0006-291X(87)90611-5

34. Popowicz AM, Dash PF (1988) SIGSEQ: a computer program for predicting signal sequence cleavage sites. Bioinformatics 4:405–406. https://doi.org/10.1093/bioinformatics/4.3.405
35. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276–277. https://doi.org/10.1016/S0168-9525(00)02024-2
36. Hiller K, Grote A, Scheer M et al (2004) PrediSi: prediction of signal peptides and their cleavage positions. Nucleic Acids Res 32:W375–W379. https://doi.org/10.1093/nar/gkh378
37. Chou K-C (2001) Prediction of protein signal sequences and their cleavage sites. Proteins Struct Funct Bioinform 42:136–139. https://doi.org/10.1002/1097-0134(20010101)42:1<136::AID-PROT130>3.0.CO;2-F
38. Nielsen H, Brunak S, Engelbrecht J, von Heijne G (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 10:1–6. https://doi.org/10.1093/protein/10.1.1
39. Chou K-C (2001) Using subsite coupling to predict signal peptides. Protein Eng Des Sel 14:75–79. https://doi.org/10.1093/protein/14.2.75
40. Chou K-C (2001) Prediction of signal peptides using scaled window. Peptides 22:1973–1979. https://doi.org/10.1016/S0196-9781(01)00540-X
41. Chou K-C, Shen H-B (2007) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Commun 357:633–640. https://doi.org/10.1016/j.bbrc.2007.03.162
42. Chou K-C (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins Struct Funct Bioinform 43:246–255. https://doi.org/10.1002/prot.1035
43. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. Nucleic Acids Res 10:2997–3011. https://doi.org/10.1093/nar/10.9.2997
44. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation. In: Rumelhart DE, McClelland JL (eds) Parallel distributed processing: explorations in the microstructure of cognition, vol 1. Foundations. MIT Press, Cambridge, pp 318–362
45. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536. https://doi.org/10.1038/323533a0
46. Kurenkov A (2015) A "Brief" history of neural nets and deep learning. In: Andrey Kurenkovs Web World. http://www.andreykurenkov.com/writing/ai/a-brief-history-of-neural-nets-and-deep-learning/. Accessed 27 Dec 2018
47. Baldi P, Brunak S (2001) Bioinformatics: the machine learning approach, 2nd edn. The MIT Press, Boston
48. Brunak S, Engelbrecht J, Knudsen S (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence. J Mol Biol 220:49–65. https://doi.org/10.1016/0022-2836(91)90380-O
49. Ladunga I, Czakó F, Csabai I, Geszti T (1991) Improving signal peptide prediction accuracy by simulated neural network. Bioinformatics 7:485–487. https://doi.org/10.1093/bioinformatics/7.4.485
50. Hertz JA, Krogh AS, Palmer RG (1991) Introduction to the theory of neural computation. Westview Press, Redwood City, Calif
51. Arrigo P, Giuliano F, Scalia F et al (1991) Identification of a new motif on nucleic acid sequence data using Kohonen's self-organizing map. Bioinformatics 7:353–357. https://doi.org/10.1093/bioinformatics/7.3.353
52. Bulmer M (1988) Codon usage and intragenic position. J Theor Biol 133:67–71. https://doi.org/10.1016/S0022-5193(88)80024-9
53. Schneider G, Wrede P (1993) Development of artificial neural filters for pattern recognition in protein sequences. J Mol Evol 36:586–595. https://doi.org/10.1007/BF00556363
54. Schneider G, Wrede P (1994) The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. Biophys J 66:335–344. https://doi.org/10.1016/S0006-3495(94)80782-9
55. Wrede P, Landt O, Klages S et al (1998) Peptide design aided by neural networks: biological activity of artificial signal peptidase I cleavage sites. Biochemistry 37:3588–3593. https://doi.org/10.1021/bi9726032
56. Darius F, Rojas R (1994) "Simulated molecular evolution" or computer-generated artifacts? Biophys J 67:2120–2122. https://doi.org/10.1016/S0006-3495(94)80695-2
57. Nielsen H, Engelbrecht J, Brunak S, Heijne GV (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Int J Neural Syst 08:581–599. https://doi.org/10.1142/S0129065797000537
58. Nielsen H, Krogh A (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. Proc Int Conf Intell Syst Mol Biol 6:122–130
59. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795. https://doi.org/10.1016/j.jmb.2004.05.028
60. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8:785–786. https://doi.org/10.1038/nmeth.1701
61. Nielsen H (2017) Predicting secretory proteins with SignalP. In: Kihara D (ed) Protein function prediction. Springer, New York, pp 59–73. https://doi.org/10.1007/978-1-4939-7015-5_6
62. von Heijne G (1989) The structure of signal peptides from bacterial lipoproteins. Protein Eng 2:531–534. https://doi.org/10.1093/protein/2.7.531
63. Berks BC (1996) A common export pathway for proteins binding complex redox cofactors? Mol Microbiol 22:393–404. https://doi.org/10.1046/j.1365-2958.1996.00114.x
64. Cristóbal S, de Gier J-W, Nielsen H, von Heijne G (1999) Competition between Sec- and TAT-dependent protein translocation in *Escherichia coli*. EMBO J 18:2982–2990. https://doi.org/10.1093/emboj/18.11.2982
65. Juncker AS, Willenbrock H, von Heijne G et al (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci 12:1652–1662. https://doi.org/10.1110/ps.0303703
66. Bendtsen JD, Nielsen H, Widdick D et al (2005) Prediction of twin-arginine signal peptides. BMC Bioinform 6:167. https://doi.org/10.1186/1471-2105-6-167
67. Fariselli P, Finocchiaro G, Casadio R (2003) SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. Bioinformatics 19:2498–2499. https://doi.org/10.1093/bioinformatics/btg360
68. Hulo N, Sigrist CJA, Le Saux V et al (2004) Recent improvements to the PROSITE database. Nucleic Acids Res 32:D134–D137. https://doi.org/10.1093/nar/gkh044
69. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436. https://doi.org/10.1038/nature14539
70. Angermueller C, Pärnamaa T, Parts L, Stegle O (2016) Deep learning for computational biology. Mol Syst Biol 12:878. https://doi.org/10.15252/msb.20156651
71. Min S, Lee B, Yoon S (2017) Deep learning in bioinformatics. Brief Bioinform 18:851–869. https://doi.org/10.1093/bib/bbw068
72. Ching T, Himmelstein DS, Beaulieu-Jones BK et al (2018) Opportunities and obstacles for deep learning in

biology and medicine. J R Soc Interface. https://doi.org/10.1098/rsif.2017.0387
73. Savojardo C, Martelli PL, Fariselli P et al (2018) DeepSig: deep learning improves signal peptide detection in proteins. Bioinformatics 34:1690–1696. https://doi.org/10.1093/bioinformatics/btx818
74. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK et al (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol 37:420–423. https://doi.org/10.1038/s41587-019-0036-z
75. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
76. Finn RD, Coggill P, Eberhardt RY et al (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44:D279–D285. https://doi.org/10.1093/nar/gkv1344
77. Chou MM, Kendall DA (1990) Polymeric sequences reveal a functional interrelationship between hydrophobicity and length of signal peptides. J Biol Chem 265:2873–2880
78. Nilsson I, Whitley P, von Heijne G (1994) The COOH-terminal ends of internal signal and signal-anchor sequences are positioned differently in the ER translocase. J Cell Biol 126:1127–1132. https://doi.org/10.1083/jcb.126.5.1127
79. Zhang Z, Wood WI (2003) A profile hidden Markov model for signal peptides generated by HMMER. Bioinformatics 19:307–308. https://doi.org/10.1093/bioinformatics/19.2.307
80. Käll L, Krogh A, Sonnhammer ELL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338:1027–1036. https://doi.org/10.1016/j.jmb.2004.03.016
81. Bagos PG, Nikolaou EP, Liakopoulos TD, Tsirigos KD (2010) Combined prediction of Tat and Sec signal peptides with hidden Markov models. Bioinformatics 26:2811–2817. https://doi.org/10.1093/bioinformatics/btq530
82. Bagos PG, Tsirigos KD, Liakopoulos TD, Hamodrakas SJ (2008) Prediction of lipoprotein signal peptides in Gram-positive bacteria with a Hidden Markov Model. J Proteome Res 7:5082–5093. https://doi.org/10.1021/pr800162c
83. Bagos PG, Tsirigos KD, Plessas SK et al (2009) Prediction of signal peptides in archaea. Protein Eng Des Sel 22:27–35. https://doi.org/10.1093/protein/gzn064
84. Vert J-P (2002) Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings. Biocomputing 2002. World Scientific Publishing, Kauai, pp 649–660
85. Cai Y-D, Lin S, Chou K-C (2003) Support vector machines for prediction of protein signal sequences and their cleavage sites. Peptides 24:159–161. https://doi.org/10.1016/S0196-9781(02)00289-9
86. Wang M, Yang J, Chou K-C (2005) Using string kernel to predict signal peptide cleavage site based on subsite coupling model. Amino Acids 28:395–402. https://doi.org/10.1007/s00726-005-0189-6
87. Nugent T, Jones DT (2009) Transmembrane protein topology prediction using support vector machines. BMC Bioinform 10:159. https://doi.org/10.1186/1471-2105-10-159
88. Jones DT (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics 23:538–544. https://doi.org/10.1093/bioinformatics/btl677
89. Shen H-B, Chou K-C (2007) Signal-3L: a 3-layer approach for predicting signal peptides. Biochem Biophys Res Commun 363:297–303. https://doi.org/10.1016/j.bbrc.2007.08.140
90. Frank K, Sippl MJ (2008) High-performance signal peptide prediction based on sequence alignment techniques. Bioinformatics 24:2172–2176. https://doi.org/10.1093/bioinformatics/btn422
91. Altschul SF, Madden TL, Schaffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389. https://doi.org/10.1093/nar/25.17.3389
92. Zhang Y-Z, Shen H-B (2017) Signal-3L 2.0: A hierarchical mixture model for enhancing protein signal peptide prediction by incorporating residue-domain cross-level features. J Chem Inf Model. https://doi.org/10.1021/acs.jcim.6b00484
93. Marchler-Bauer A, Derbyshire MK, Gonzales NR et al (2015) CDD: NCBI's conserved domain database. Nucleic Acids Res 43:D222–D226. https://doi.org/10.1093/nar/gku1221
94. Fukasawa Y, Leung RK, Tsui SK, Horton P (2014) Plus ça change—evolutionary sequence divergence predicts protein subcellular localization signals. BMC Genomics 15:46. https://doi.org/10.1186/1471-2164-15-46
95. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580. https://doi.org/10.1006/jmbi.2000.4315
96. von Heijne G (2007) The membrane protein universe: what's out there and why bother? J Intern Med 261:543–557. https://doi.org/10.1111/j.1365-2796.2007.01792.x
97. Tsirigos KD, Govindarajan S, Bassot C et al (2018) Topology of membrane proteins—predictions, limitations and variations. Curr Opin Struct Biol 50:9–17. https://doi.org/10.1016/j.sbi.2017.10.003
98. Lao DM, Arai M, Ikeda M, Shimizu T (2002) The presence of signal peptide significantly affects transmembrane topology prediction. Bioinformatics 18:1562–1566. https://doi.org/10.1093/bioinformatics/18.12.1562
99. Lao DM, Okuno T, Shimizu T (2002) Evaluating transmembrane topology prediction methods for the effect of signal peptide in topology prediction. Silico Biol 2:485–494
100. Argos P, Rao JKM, Hargrave PA (1982) Structural prediction of membrane-bound proteins. Eur J Biochem 128:565–575. https://doi.org/10.1111/j.1432-1033.1982.tb07002.x
101. Klein P, Kanehisa M, DeLisi C (1985) The detection and classification of membrane-spanning proteins. Biochim Biophys Acta BBA 815:468–476. https://doi.org/10.1016/0005-2736(85)90375-X
102. von Heijne G (1986) The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. EMBO J 5:3021–3027. https://doi.org/10.1002/j.1460-2075.1986.tb04601.x
103. von Heijne G, Gavel Y (1988) Topogenic signals in integral membrane proteins. Eur J Biochem 174:671–678. https://doi.org/10.1111/j.1432-1033.1988.tb14150.x
104. von Heijne G (1992) Membrane protein structure prediction: hydrophobicity analysis and the positive-inside rule. J Mol Biol 225:487–494. https://doi.org/10.1016/0022-2836(92)90934-C
105. Claros MG, von Heijne G (1994) TopPred II: an improved software for membrane protein structure predictions. Bioinformatics 10:685–686. https://doi.org/10.1093/bioinformatics/10.6.685
106. Jones DT, Taylor WR, Thornton JM (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry 33:3038–3049. https://doi.org/10.1021/bi00176a037
107. Jones DT (1998) Do transmembrane protein superfolds exist? FEBS Lett 423:281–285. https://doi.org/10.1016/S0014-5793(98)00095-7
108. Pasquier C, Promponas VJ, Palaios GA et al (1999) A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the SwissProt database: the PRED-TMR algorithm. Protein Eng 12:381–385. https://doi.org/10.1093/protein/12.5.381

109. Bernsel A, Viklund H, Falk J et al (2008) Prediction of membrane-protein topology from first principles. Proc Natl Acad Sci 105:7177–7181. https://doi.org/10.1073/pnas.0711151105
110. Peters C, Tsirigos KD, Shu N, Elofsson A (2016) Improved topology prediction using the terminal hydrophobic helices rule. Bioinformatics 32:1158–1162. https://doi.org/10.1093/bioinformatics/btv709
111. Sonnhammer ELL, von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 6:175–182
112. Tusnády GE, Simon I (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol 283:489–506. https://doi.org/10.1006/jmbi.1998.2107
113. Tusnády GE, Simon I (2001) The HMMTOP transmembrane topology prediction server. Bioinformatics 17:849–850. https://doi.org/10.1093/bioinformatics/17.9.849
114. Käll L, Krogh A, Sonnhammer ELL (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. Bioinformatics 21:i251–i257. https://doi.org/10.1093/bioinformatics/bti1014
115. Reynolds SM, Käll L, Riffle ME et al (2008) Transmembrane topology and signal peptide prediction using dynamic bayesian networks. PLoS Comput Biol 4:e1000213. https://doi.org/10.1371/journal.pcbi.1000213
116. Viklund H, Bernsel A, Skwark M, Elofsson A (2008) SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. Bioinformatics 24:2928–2929. https://doi.org/10.1093/bioinformatics/btn550
117. Rost B, Sander C (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci USA 90:7558–7562. https://doi.org/10.1073/pnas.90.16.7558
118. Persson B, Argos P (1994) Prediction of transmembrane segments in proteins utilising multiple sequence alignments. J Mol Biol 237:182–192. https://doi.org/10.1006/jmbi.1994.1220
119. Rost B, Sander C, Casadio R, Fariselli P (1995) Transmembrane helices predicted at 95% accuracy. Protein Sci 4:521–533. https://doi.org/10.1002/pro.5560040318
120. Viklund H, Elofsson A (2004) Best α-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. Protein Sci Publ Protein Soc 13:1908–1917. https://doi.org/10.1110/ps.04625404
121. Viklund H, Elofsson A (2008) OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. Bioinformatics 24:1662–1668. https://doi.org/10.1093/bioinformatics/btn221
122. Promponas VJ, Palaios GA, Pasquier CM et al (1998) CoPreTHi: a Web tool which combines transmembrane protein segment prediction methods. Silico Biol 1:0014
123. Bernsel A, Viklund H, Hennerdal A, Elofsson A (2009) TOPCONS: consensus prediction of membrane protein topology. Nucleic Acids Res 37:W465–W468. https://doi.org/10.1093/nar/gkp363
124. Klammer M, Messina D, Schmitt T, Sonnhammer E (2009) MetaTM—a consensus method for transmembrane protein topology prediction. BMC Bioinform 10:314. https://doi.org/10.1186/1471-2105-10-314
125. Dobson L, Reményi I, Tusnády GE (2015) CCTOP: a consensus constrained TOPology prediction web server. Nucleic Acids Res 43:W408–W412. https://doi.org/10.1093/nar/gkv451
126. Tsirigos KD, Peters C, Shu N et al (2015) The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res 43:W401–W407. https://doi.org/10.1093/nar/gkv485
127. Zhai Y, Saier MH (2002) The β-barrel finder (BBF) program, allowing identification of outer membrane β-barrel proteins encoded within prokaryotic genomes. Protein Sci Publ Protein Soc 11:2196–2207. https://doi.org/10.1110/ps.0209002
128. Wimley WC (2002) Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. Protein Sci Publ Protein Soc 11:301–312. https://doi.org/10.1110/ps.29402
129. Remmert M, Linke D, Lupas AN, Söding J (2009) HHomp—prediction and classification of outer membrane proteins. Nucleic Acids Res. https://doi.org/10.1093/nar/gkp325
130. Martelli PL, Fariselli P, Krogh A, Casadio R (2002) A sequence-profile-based HMM for predicting and discriminating β barrel membrane proteins. Bioinformatics 18:S46–S53. https://doi.org/10.1093/bioinformatics/18.suppl_1.S46
131. Bagos P, Liakopoulos T, Spyropoulos I, Hamodrakas S (2004) A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins. BMC Bioinform 5:29. https://doi.org/10.1186/1471-2105-5-29
132. Bigelow HR, Petrey DS, Liu J et al (2004) Predicting transmembrane beta-barrels in proteomes. Nucleic Acids Res 32:2566–2577. https://doi.org/10.1093/nar/gkh580
133. Savojardo C, Fariselli P, Casadio R (2013) BETAWARE: a machine-learning tool to detect and predict transmembrane beta-barrel proteins in prokaryotes. Bioinformatics 29:504–505. https://doi.org/10.1093/bioinformatics/bts728
134. Tsirigos KD, Elofsson A, Bagos PG (2016) PRED-TMBB2: improved topology prediction and detection of beta-barrel outer membrane proteins. Bioinformatics 32:i665–i671. https://doi.org/10.1093/bioinformatics/btw444
135. Hayat S, Elofsson A (2012) BOCTOPUS: improved topology prediction of transmembrane β barrel proteins. Bioinformatics 28:516–522. https://doi.org/10.1093/bioinformatics/btr710
136. Hayat S, Peters C, Shu N et al (2016) Inclusion of dyad-repeat pattern improves topology prediction of transmembrane β-barrel proteins. Bioinformatics 32:1571–1573. https://doi.org/10.1093/bioinformatics/btw025
137. Jacoboni I, Martelli PL, Fariselli P et al (2001) Prediction of the transmembrane regions of β-barrel membrane proteins with a neural network-based predictor. Protein Sci Publ Protein Soc 10:779–787. https://doi.org/10.1110/ps.37201
138. Gromiha MM, Ahmad S, Suwa M (2004) Neural network-based prediction of transmembrane β-strand segments in outer membrane proteins. J Comput Chem 25:762–767. https://doi.org/10.1002/jcc.10386
139. Ou Y-Y, Gromiha MM, Chen S-A, Suwa M (2008) TMBETA-DISC-RBF: discrimination of β-barrel membrane proteins using RBF networks and PSSM profiles. Comput Biol Chem 32:227–231. https://doi.org/10.1016/j.compbiolchem.2008.03.002
140. Horton P, Nakai K (1996) A probabilistic classification system for predicting the cellular localization sites of proteins. Proc Int Conf Intell Syst Mol Biol 4:109–115
141. Horton P, Nakai K (1997) Better prediction of protein cellular localization sites with the k nearest neighbors classifier. Proc Int Conf Intell Syst Mol Biol 5:147–152
142. Nakashima H, Nishikawa K (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J Mol Biol 238:54–61. https://doi.org/10.1006/jmbi.1994.1267
143. Cedano J, Aloy P, Pérez-Pons JA, Querol E (1997) Relation between amino acid composition and cellular location of proteins. J Mol Biol 266:594–600. https://doi.org/10.1006/jmbi.1996.0804
144. Chou K-C, Elrod DW (1998) Using discriminant function for prediction of subcellular location of prokaryotic proteins. Biochem

Biophys Res Commun 252:63–68. https://doi.org/10.1006/bbrc.1998.9498
145. Chou K-C, Elrod DW (1999) Prediction of membrane protein types and subcellular locations. Proteins Struct Funct Bioinform 34:137–153. https://doi.org/10.1002/(SICI)1097-0134(19990101)34:1<137::AID-PROT11>3.0.CO;2-O
146. Chou K-C, Elrod DW (1999) Protein subcellular location prediction. Protein Eng Des Sel 12:107–118. https://doi.org/10.1093/protein/12.2.107
147. Reinhardt A, Hubbard T (1998) Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 26:2230–2236. https://doi.org/10.1093/nar/26.9.2230
148. Yuan Z (1999) Prediction of protein subcellular locations using Markov chain models. FEBS Lett 451:23–26. https://doi.org/10.1016/S0014-5793(99)00506-2
149. Hua S, Sun Z (2001) Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17:721–728. https://doi.org/10.1093/bioinformatics/17.8.721
150. Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins Struct Funct Bioinform 9:56–68. https://doi.org/10.1002/prot.340090107
151. Hobohm U, Scharf M, Schneider R, Sander C (1992) Selection of representative protein data sets. Protein Sci 1:409–417. https://doi.org/10.1002/pro.5560010313
152. Nielsen H, Engelbrecht J, von Heijne G, Brunak S (1996) Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site. Proteins Struct Funct Bioinform 24:165–177. https://doi.org/10.1002/(SICI)1097-0134(199602)24:2<165::AID-PROT4>3.0.CO;2-I
153. Nair R, Rost B (2005) Mimicking cellular sorting improves prediction of subcellular localization. J Mol Biol 348:85–100. https://doi.org/10.1016/j.jmb.2005.02.025
154. Yu C-S, Chen Y-C, Lu C-H, Hwang J-K (2006) Prediction of protein subcellular localization. Proteins 64:643–651. https://doi.org/10.1002/prot.21018
155. Pierleoni A, Martelli PL, Fariselli P, Casadio R (2006) BaCelLo: a balanced subcellular localization predictor. Bioinformatics 22:e408–e416. https://doi.org/10.1093/bioinformatics/btl222
156. Xie D, Li A, Wang M et al (2005) LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. Nucleic Acids Res 33:W105–W110. https://doi.org/10.1093/nar/gki359
157. Bhasin M, Raghava GPS (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res 32:W414–W419. https://doi.org/10.1093/nar/gkh350
158. Andrade MA, O'Donoghue SI, Rost B (1998) Adaptation of protein surfaces to subcellular location. J Mol Biol 276:517–525. https://doi.org/10.1006/jmbi.1997.1498
159. Nair R, Rost B (2002) Sequence conserved for subcellular localization. Protein Sci 11:2836–2847. https://doi.org/10.1110/ps.0207402
160. Imai K, Nakai K (2010) Prediction of subcellular locations of proteins: where to proceed? Proteomics 10:3970–3983. https://doi.org/10.1002/pmic.201000274
161. Blum T, Briesemeister S, Kohlbacher O (2009) MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. BMC Bioinform 10:274. https://doi.org/10.1186/1471-2105-10-274
162. Horton P, Park K-J, Obayashi T et al (2007) WoLF PSORT: protein localization predictor. Nucleic Acids Res 35:W585–W587. https://doi.org/10.1093/nar/gkm259
163. Goldberg T, Hecht M, Hamp T et al (2014) LocTree3 prediction of localization. Nucleic Acids Res 42:W350–W355. https://doi.org/10.1093/nar/gku396
164. Goldberg T, Hamp T, Rost B (2012) LocTree2 predicts localization for all domains of life. Bioinformatics 28:i458–i465. https://doi.org/10.1093/bioinformatics/bts390
165. Yu NY, Wagner JR, Laird MR et al (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 26:1608–1615. https://doi.org/10.1093/bioinformatics/btq249
166. Marcotte EM, Xenarios I, van der Bliek AM, Eisenberg D (2000) Localizing proteins in the cell from their phylogenetic profiles. Proc Natl Acad Sci 97:12115–12120. https://doi.org/10.1073/pnas.220399497
167. Scott MS, Thomas DY, Hallett MT (2004) Predicting subcellular localization via protein motif co-occurrence. Genome Res 14:1957–1966. https://doi.org/10.1101/gr.2650004
168. Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25–29. https://doi.org/10.1038/75556
169. Wan S, Mak M-W, Kung S-Y (2012) mGOASVM: multi-label protein subcellular localization based on gene ontology and support vector machines. BMC Bioinform 13:290. https://doi.org/10.1186/1471-2105-13-290
170. Wan S, Mak M-W, Kung S-Y (2013) GOASVM: a subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition. J Theor Biol 323:40–48. https://doi.org/10.1016/j.jtbi.2013.01.012
171. Chou K-C, Wu Z-C, Xiao X (2011) iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. PLoS ONE 6:e18258. https://doi.org/10.1371/journal.pone.0018258
172. Chou K-C, Wu Z-C, Xiao X (2012) iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol BioSyst 8:629–641. https://doi.org/10.1039/C1MB05420A
173. Lin W-Z, Fang J-A, Xiao X, Chou K-C (2013) iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. Mol BioSyst 9:634–644. https://doi.org/10.1039/C3MB25466F
174. Wu Z-C, Xiao X, Chou K-C (2011) iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. Mol BioSyst 7:3287–3297. https://doi.org/10.1039/C1MB05232B
175. Wu Z-C, Xiao X, Chou K-C (2012) iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex Gram-positive bacterial proteins. Protein Pept Lett 19:4–14. https://doi.org/10.2174/092986612798472839
176. Xiao X, Wu Z-C, Chou K-C (2011) iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. J Theor Biol 284:42–51. https://doi.org/10.1016/j.jtbi.2011.06.005
177. Xiao X, Wu Z-C, Chou K-C (2011) A multi-label classifier for predicting the subcellular localization of Gram-negative bacterial proteins with both single and multiple sites. PLoS ONE 6:e20592. https://doi.org/10.1371/journal.pone.0020592
178. Lu Z, Szafron D, Greiner R et al (2004) Predicting subcellular localization of proteins using machine-learned classifiers. Bioinformatics 20:547–556. https://doi.org/10.1093/bioinformatics/btg447
179. Shatkay H, Höglund A, Brady S et al (2007) SherLoc: high-accuracy prediction of protein subcellular localization by integrating

text and protein sequence data. Bioinformatics 23:1410–1417. https://doi.org/10.1093/bioinformatics/btm115
180. Briesemeister S, Blum T, Brady S et al (2009) SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins. J Proteome Res 8:5363–5366. https://doi.org/10.1021/pr900665y
181. Gardy JL, Spencer C, Wang K et al (2003) PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. Nucleic Acids Res 31:3613–3617. https://doi.org/10.1093/nar/gkg602
182. Gardy JL, Laird MR, Chen F et al (2005) PSORTb vol 2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics 21:617–623. https://doi.org/10.1093/bioinformatics/bti057
183. Höglund A, Dönnes P, Blum T et al (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. Bioinformatics 22:1158–1165. https://doi.org/10.1093/bioinformatics/btl002
184. Briesemeister S, Rahnenfuhrer J, Kohlbacher O (2010) YLoc—an interpretable web server for predicting subcellular localization. Nucleic Acids Res 38:W497–W502. https://doi.org/10.1093/nar/gkq477
185. Briesemeister S, Rahnenführer J, Kohlbacher O (2010) Going from where to why—interpretable prediction of protein subcellular localization. Bioinformatics 26:1232–1238. https://doi.org/10.1093/bioinformatics/btq115
186. Almagro Armenteros JJ, Sønderby CK, Nielsen H, Winther O (2017) DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics 33:3387–3395. https://doi.org/10.1093/bioinformatics/btx431